A Literature Summary:
Reduce Clever-Hans Effects in Neural Networks

Jim Neuendorf
Explainable AI for Decision Making, SS 2023

Abstract

With the rise of popularity of artificial intelligence (AI) in daily applications, the need for the explanation of a model's prediction has also become more important. Some of the current research focuses on neural networks (NN) that produce the correct output for the wrong reasons. This effect is called the Clever Hans (CH). The papers summarized in this work contribute methods related to explanation for reducing these effects and thus making models more robust.

1. Introduction

AI is used in all kinds of applications across many domains. In certain situations, for example medical assistance, decision errors are intolerable. Since AI models typically do not achieve 100% accuracy on complex tasks, explainable AI (XAI) is required to establish trust, fairness and transparency by providing answers to *wh*-questions, e.g. "**Wh**y did the model predict something?" Sometimes, explainability is even mandatory by law such as in the European Union with its *right to explanation*. Consequently, XAI techniques must applicable to models processing all imaginable types of data, i.e. images, audio, video, text etc. [1]. Furthermore, foundation models are becoming popular so fighting CH effects is a timely concern [2].

2. Overview

In order to detect CH effects, XAI can be a helpful tool. Therefore, we first look at "methods to make explanations [...] more robust against attacks that manipulate the input" [3]. We then discuss methods for reducing CH effects using XAI that either (1) retrain the model using modified data or an adjusted objective or (2) change the model structurally, e.g. by inserting layers. The latter is most common among the following and called *post-hoc* because the model regarded is already trained [1].

3. Robust Explanations

As mentioned in Section 2 explanations can be made more robust using three relatively simple techniques. These are derived from the finding that the maximum possible explanation change is bounded by the Frobenius norm of the Hessian $\|H\|_F$ [3]. Hence, the presented approaches try to minimize $\|H\|_F$ – within the context of supervised learning.

3.1. Curvature Minimization

An intuitive solution is to penalize the F-Norm by adding a term to the loss function. Because “calculating the Frobenius norm of the Hessian es expensive, [...] [they] propose to estimate the F-norm stochastically” [3] with an expectation value using Monte-Carlo sampling.

Note that the impact of the penalization term can be controlled by a hyperparameter.

3.2. Weight Decay

Because $\|H\|_F$ also depends on the weights of the NN, weight decay can also be used to robustify the explanation. This method is a regularization aiming for small weight values. In this case using the Frobenius norm – other regularizations such as L^2 could work as well. Weight decay is a well-known technique for improving model generalization, but in the context of XAI it is new.

It is notable that a technique that improves model generalization simultaneously helps enhancing a model’s explanation robustness.

3.3. Smooth Activation Functions

As a third approach activation functions with smaller maximum values of the first and second derivatives result in smaller values for $\|H\|_F$. The Softplus non-linearity is used as an example for a smooth function. However, effects on training of the model are not presented (e.g. in constrast to ReLU).

3.4. Results

The three presented techniques could be validated independently using a convolutional neural network (CNN) on the CIFAR-10 dataset. Even a small reduction of the curvature led to a significant improvement of the explanation robustness. However, the model accuracy decreases the more the robustness increases. Thus, there is a trade-off between the two effects.

4. Reducing Clever-Hans Effects

This section contains attempts to reduce CH effects in both supervised and unsupervised settings using post-hoc as well as fine-tuning approaches.

4.1. Bagging in Anomaly Detection

Research on “whether Clever Hans also occurs in unsupervised learning, and in which form, has received so far alsmost no attention.” [4] So far only model-specific approaches have been proposed, i.e. no general technique that applies to all anomaly detection models. Anomaly detection approaches can be categorized into *density-*, *reconstruction-*, and *boundary-based.*

Therefore, the researchers “introduce a common XAI framework that is applicable to a broad range of anomaly detection models” [4]: They use a NN architecture with three layers for explanation extraction using Deep Tayler Decomposition (a mechanism similar to Layer-wise Relevance Propagation (LRP)). For each of the three layers *feature extraction*, *distance*, and *pooling* they define (1) specific calculations and (2) certain propagation rules – both depending on the type of model[1].

For example, the distance (forward pass) of a reconstruction-based model is the Mean Squared Error (MSE), its "outlier score", $o(x) = \|r(x) - x\|^2$. For the attribution (backward pass) in the distance layer the mean μ_{jk} is actually used as opposed to being a constant for the other model types.

4.1.1. Evaluation

Using the datasets MNIST-C and MVTec, both coming with anomaly-ground-truth data, and a Clever-Hans score[2], they find that the different "models are affected by the problem in different ways" [4]. Thus, they "hypothesize that [CH effects] are inherent to the structre of the anomaly detection models rather than […] the trainig data." [4] This hypothesis is followed by a reasoning about why this is the case.

Sticking with the autoencoder example, they argue that the CH effect is caused by samples whose reconstruction is far away from the input distribution. Thus, an outlier is detected due to features that do not relate to the input distribution which means the outlier detection is unrelated to the actual anomaly.

4.1.2. Solution

Because the source of CH is not in the data but in in the models themselves, the idea is to allow "multiple anomaly models to mutually cancel their individual structural weaknesses" [4]. Thus, a bagging approach of outlier scores is proposed:

$$o_{\text{Bag}}(x) = \frac{1}{3}\Big(o_{\text{KDE}}(x) + o_{\text{Auto}}(x) + o_{\text{Deep}}(x)\Big)$$

where KDE is density-, Auto is boundary-, and Deep is boundary-based.

"The bagged model ranks first among all four models […] although relatively far from the ground-truth" [4]. So there is still a lot of room for improvement with techniques going "beyond simple bagging […] and structurally less rigid models" [4].

4.2. Pruning in Deep Models

CH effects can remain undetected even if the user's explanation agrees with the one from XAI. In this case, there is "neither prior knowledge about the spurious feature, nor data containing it" [2]: We receive the trained model which "is to be robustified post-hoc with limited data" [2] without CH features, apply the soft-pruning, and should be able to deploy the model.

The proposed pruning method is called *Explanation-Guided Exposure Minimization* (EGEM). They arrive at a practical formulation which comprises (1) explanations of the refined model to be close to those of the original one and (2) a penalty used for exposure minimization. The pruning strength depends on each neuron's activation frequency and magnitude.

A second pruning approach *PCA-EGEM* is introduced for more effective pruning: In PCA space the features are more disentangled so pruning should be more effective.

[1] The propagation rules (backward pass) are model-dependent because the calculation (forward pass) also is.

[2] The score measures the mismatch between the detection and explanation accuracy.

4.2.1. Evaluation

In their work the scientists could validate approve the effectiveness of the pruning method using datasets were the spurious artifacts were known: MNIST with manual poisoning, and ImageNet and ISIC were previous work has identified (potential) CH correlations.

Afterwards, the approach was tested in an exploratory manner on the CelebA dataset. Using PCA-EGEM it could be observed that "a model bias [...] has been mitigated" and it "enables the retrieval of a more diverse set of positive instances from a large heterogeneous dataset" [2].

4.3. Class Artifact Compensation in Deep Models

Another unlearning approach using XAI called *Class Artifact Compensation* (ClArC) is provided by [5]. This work focuses on automating the process of CH removal because with datasets getting larger, manual inspection/curation becomes infeasible.

4.3.1. Large-scale Analysis

For semi-automated discovery a technique called *Spectral Relevance Analysis* (SpRAy) is used which aims to bridge local and global XAI by inspecting large sets of local explanations. From attribution maps it computes a spectrum of eigenvalues via spectral clusters. The eigenvalue spectrum can be used "for ranking [...] analyzed classes w.r.t. their potential for exhibiting CH phenomena" [5].

In addition to enhancing SpRAy's cluster visualization and labeling by making use of an intermediate result, they propose using Fisher Discriminant Analysis[3] (FDA) to rank class-wise clusterings by their separability τ, especially when there are many classes resulting in many clusters [5]. Large values of τ indicate artifiact candidates.

4.3.2. Unlearning with Augmentative Class Artifact Compensation

Here (A-ClArC), the training data is augmented in a way that the trained classifier becomes insensitive to an artifact. This is achieved by adding an artifact found with SpRAy to some samples of all other classes using a *forward artifact model.*

When the classifier is trained with the augmented data, it can no longer rely on the artifact as an "easy shortcut" for classification but must learn other strategies.

4.3.3. Unlearning with Projective Class Artifact Compensation

In constrast to the before-mentioned method that requires retraining the model, the second proposed approach (P-ClArC) suppresses CH artifacts[4]. This happens "without retraining by incorporating a *backward artifact model* [...] directly into the prediction model" [5]. While the idea is basically the same as for A-ClArC, it works by projecting the data points to a position of the decision boundary to which the estimated articat direction is normal. Therefore, the boundary ignores and thus suppresses the artifact while leaving the output unchanged for clean samples.

[3] FDA maximizes between-class and minimizes within-class scatter. Other separation algorithms possible.

[4] It does not perform true unlearning in the same way as A-ClArC.

4.3.4. Evaluation

The paper contains a very thorough evaluation of the extended SpRAy and ClArC methods on six datasets. What they found is that their extended SpRAy (1) picks up most artifacts when there are any, even though their importance varies between models and classes, and (2) still requires human judgement for the final decision on CH candidates.

Concerning their proposed methods they conclude that both A-ClArC and P-ClArC perform very well – in input as well as latent space. Two mentioned limitations are that A-ClArC is time-consuming and "P-ClArC will not lead to an increased generalization performance, since the model never has a chance to adapt its weights [...] and correct its faulty prediction" [5].

Hence, common datasets like ImageNet can be "un-Hansed" to provide a more unbiased basis for foundation models.

4.4. Post-hoc Explanations for Unknown Spurious Correlation

The post-hoc approaches discussed so far all relied on XAI and seemed very promising. In the last paper of this summary, the "post-hoc explanation methods tested are ineffective when the spurious artifact is unknwon at test-time" [6]. The used explanations methods are (1) feature attribution, (2) concept activation, and (3) training point ranking.

In order to compare these methods, they define a *spurious score*[5] which quantifies "the strength of a model's dependence on a training signal" [3].

For their evaluation, they use three measures comparing different combinations of explanations[6] with three similarity functions[7]. Their findings range from (1) struggling "to detect blur signals even when known" over (2) "concept rankings can help detect reliance on the visible signals but not non-visible" to (3) "all methods struggle to reliably indicate that spurious models are reliant on the blur signal" [3].

5. Summary

With the last paper dulling the results outlined previously, it will be interesting to see how the big picture evolves, e.g. how P-ClArC performs across different models and tasks when evaluated in the spurious-score settings. Additionally, further research on how the structure of a model affects explanations and their robustness in supervised learning could impact the development of future foundation models[8].

References

[1] P. Gohel, P. Singh, and M. Mohanty, "Explainable AI: Current status and future directions," 2021.

[2] L. Linhardt, K.-R. Müller, and G. Montavon, "Preemptively pruning clever-hans strategies in deep neural networks," 2023.

[5] In other words, it "is the probability that the model assigns the input to the spurious aligned class if the spurious signal is added to the input" [3].

[6] (1)

[7] One for each attribution method.

[8] For example, DINOv2 has amazing emergent properties concerning the latent-space structure.

[3] A.-K. Dombrowski, C. J. Anders, K.-R. Müller, and P. Kessel, "Towards robust explanations for deep neural networks," 2020.

[4] J. Kauffmann, L. Ruff, G. Montavon, and K.-R. Müller, "The clever hans effect in anomaly detection," 2020.

[5] C. J. Anders, L. Weber, et al., "Finding and removing clever hans: Using explanation methods to debug and improve deep models," 2020.

[6] J. Adebayo, M. Muelly, H. Abelson, and B. Kim, Post hoc explanations may be ineffective for detecting unknown spurious correlation. Presented at Internation Conf. Learn. Representations, 2022.